Model Design and Implementation of Enterprise Credit Information Based On Data Mining

Qingyuan Xu
Digital China Information Service Company LTD.
Beijing, 100085, China,
13910578378@139.com

Qingyue Xu
Beijing Union University, Beijing, 100101, China,
lytxuqingyue@buu.edu.cn

Abstract

Smart city construction emphasizes building an effective whole social credit system, promoting the construction of government integrity, business integrity, social integrity and public confidence in the judiciary. For credit risk has become an important assessment. Meanwhile, administration credit system is one of three major data. It proposes unified credit discipline and warning regulatory purposes, led by the government and its main functions, taken governmental data as the main basis. Accordingly, the paper constructs Corporate Dishonest Credit Executed (CDCE) Risk Assessment Model, based on governmental data. The model uses a set of urban enterprise data, selecting the explanatory variables from five aspects, administrative punishment, innovation, credit information, credit situation, and social responsibility, to screen CDCE Logit regression, to filter out and find out those variables which are significantly predicted effects for CDCE risk. And then construct a Logit regression model with the above selected variables. The experimental results and comparison of practical applications in China, we found that the model promises to higher business risk identification accuracy for CDCE. The model has a higher applied value and developmental prospects.

Keywords: *Smart city, Dishonest to enforcement risk, Dishonest enterprise to enforcement, Logit regression, risk*

1. Introduction

In the modern society, with market economy, honesty or integrity becomes the important prerequisite for human's interaction and the interaction between individual and enterprise. Economic performance and social management are being progressively credit standards, constructing a credit environment of social order, social credit environment system and economic exchange system. And with the development of the credit, credit risk is emerging objectively. Credit risk is the objective existence of the activity. It refers to one credit activity due to non-compliance and the risk of damage to the other party. Objectively, Credit risk is closely related to the specific credit relations and the normal credit order. The credit activity is more specific, credit risk is lower; Otherwise, higher. Since the modern market economy is a highly developed credit economy, credit relations across all areas of economic life. The specific credit relations and the ordered credit activity have particular importance for the normal operation of a modern market economy. So, a company's credit risk assessment, for the healthy development of enterprises in the era of credit economy, has great significance.

Meanwhile, the amount of the government controlled data has been unprecedented growth, along with the construction and development of e-government and smart city. The government data, from the management body, includes corporations, institutions and other organizations and individuals. From management business, it includes social security, industry and commerce, taxation and other social subject activities. It records the complete cycle of activity subject data from the process. From the data structure, it includes both structured information that computer handled, and the other unstructured information such as text, audio and video.

Since the body of governmental data has the characteristics of authoritative, creditability, collecting timely and complete structure, using data mining techniques has become a new direction of data mining to achieve the issuer credit rating, risk assessment, credit management. It relies mainly on government data while non-governmental data subsidiary.

With perspective, governmental data is the basis of credit data mining. It is the data of acquisition, production and management, in the process of achieving operational goals. Only the accumulation of data and effective application is to achieve the smooth development of credit business. Full view of governmental credit management needs to achieve resource sharing, through information, for the various branches of government, on the basis of business sector's data.

URI: http://hdl.handle.net/10125/41294
ISBN: 978-0-9981331-0-2

According to uniform data collection standards, various government departments obtain business data, to achieve credit information integration, application and mining through interoperability. And the credit information is cross-sectoral, cross-regional and cross-level. It also achieves governmental credit management mechanism, for government organizations' and individuals' credit, to achieve a full range of credit ratings, scoring and risk analysis. The goal is to penalizing bad, trustworthy reward objectives.

But, the punishment of Dishonest is not the goal of governmental management. But to reduce the loss brought and caused by acts of dishonesty and to find the risk of dishonesty. This requires an effective evaluation of dishonest risk. And Credit Scoring is most commonly used and the most effective way for Credit risk assessment.

Credit scoring method mainly includes multiple discriminate analysis, probit model and Logit model, etc. More and more scholars found that multiple discriminate analysis method requires independent variable applied with normal distribution. Population covariance matrix of each phase is equal, and so on. Probit model is similar to multivariate discriminant analysis. It needs to study sample data applied with normal distribution. Logit Assumptions of Logit model is more relaxed, without demand of variable meeting with normal distribution and equal covariance. The Model can directly predict the probability P of the events. As the sample data, acquired by credit risk assessment, is generally not subject to standard normal distribution, it is more suitable to use Logit regression analysis in the study of credit risk.

On the basis of obtaining the enterprise credit risk data of C City, J Province, China, and with the usage of data mining technology, this paper constructs the enterprise risk assessment method based on the Logit model. It lays the foundation for the promotion and application of corporate credit risk assessment models.

The first part of this paper describes the background of the research, The second part reviews literature of domestic and foreign enterprises credit risk assessment research. The third part is based on the Logit model to assess the dishonest for enforcement risk, It introduces the selected explanatory variable selection and model in detail, and Logit regression results analysis. The model test is completed by comparing the prediction results. The fourth part introduces the application of the model in C city of J Province in China. The fifth part obtains the conclusion, combining with the results related results, and summarizes the research contents and future work.

2. Literature review

As credit risk management is a critical important issue in banking and finance, a number of indicators and models are developed for credit risk evaluation. Initially, 5C, 5P and 5W rules are applied to credit scoring field. For example, the 5Cs rules are a set of the subjective judgment rules include character, capacity, capital, collateral and condition. However, the booming of the credit industry made it impossible to assess thousands of applicants completely manually but to automate the process. Hence various credit scoring models have emerged to help financial institutions enforcing effective credit approval. Corazza, Funari and Gusso [1] presented seven economic and financial indicators for enterprise credit modeling by using MURAME. Tsaih et al. [2] used fundamental characteristics, bank borrow relationship, personal trade history, and industry factors including business cycles, macroeconomic factors as input features and developed a probit model to evaluate enterprise credit. Similarly, Andreeva, Calabrese, and Osmetti [3] extracted profitability, leverage, coverage, liquidity, scale and non-financial information and applied generalized extreme value models to UK and Italian small businesses. Serrano-Cinca, Gutierrez-Nieto and Reyes [4] considered financial information and social impact information to credit modeling, while Fernandes and Artes [5] introduced spatial dependence into credit risk assessment to improve the performance of credit scoring. Ju, Jeon and Sohn [6] found technology-based loan default is related not only to technology-oriented attributes, such as management, technology, profitability and market ability, and firm-specific characteristics, but also to the economic situation, so a credit scoring model is proposed with various scenarios. Furthermore, Zhang et al. [7] proposes a novel credit risk evaluation approach using fundamental characteristics and sentiment indexes generated by mining the opinions of the text content in customer due diligence reports to automate the decision-making process.

As can be seen from literatures, enterprise fundamental characteristics, financial reports, social and environment indicators are involved in enterprise credit risk evaluation [1-7], and the basic rules, and several sophisticated models are designed and implemented in credit scoring [8-14]. Both the novel indicators and the state-of-art models can improve the performance of credit risk evaluation.

In recent decades, Chinese business loans grow rapidly, especially the loan demand from small and medium enterprises. However, it is critically difficult to evaluate their credit because of lacks of available data. With continuous open data in government, it provides potential data sources for enterprise credit scoring.

This paper focuses on credit model design and implementation by using enterprise credit information. Furthermore, because what we obtained the enterprise credit information cannot apply with the standard normal distribution, it is more appropriate to use logit regression for modeling.

3. Model design

3.1. Data Sources

We select a total sample of 46827 enterprises, registered in C city, credit data set is from 48 government agencies in C city.

We divide the dataset into five categories: administrative punishment, innovation consciousness, credit information, credit situation, and social responsibility, select appropriate explanatory variables, and then carry out Logit regression to the dishonest enterprise enforcement. From this, we select and find out the significant variables which can predict effectively for the credibility of the dishonest enterprise.

Then, we construct Logit regression model by using these selected variables. According to these test data, the model is used to predict. To verify the model practicability and generalization ability, we compare the consistent of the enterprise credibility with actual situation. These enterprises are predicted by model.

3.2. Variable selection

3.2.1 Dependent variable selection

The dependent variable in this paper is dishonest enterprise to enforcement in C city. Based on the list of public dishonesty executed in C city, we mark the enterprises. If the existence of the enterprise dishonesty is executed, the mark is 1, and the contrary is marked as 0.

3.2.2 Explanatory variable selection

According to the existing enterprise credit data set, weselect13 explanatory variables to build a preliminary model, then remove the explanatory variables with poor ability to explained dependent variable, finally, the rest of the explanatory variables are used to construct enterprise credit risk assessment model.

Preliminary modeling of explanatory variables, table 1

Table 1. Preliminary modeling of explanatory variables

measurefactor	variable code	variable name	definition method
administrative punishment	VAR1	number of administrative punishment	total calculation
innovation consciousness	VAR2	Whether to get high-tech enterprise identification	virtual variable. 1: yes, 0: no
	VAR3	Whether to get the award of science and technology innovation	virtual variable. 1: yes, 0: no
	VAR4	Whether the introduction of high-end talent	virtual variable. 1: yes, 0: no
credit information	VAR5	Credit line and registered capital ratio	bank credit line divided by registered capital
	VAR6	Whether to carry out normal business discount	virtual variable. 1: yes, 0: no
	VAR7	Whether there is a non normal credit	virtual variable. 1: yes, 0: no
credit situation	VAR8	Whether the implementation of enterprise credit management	virtual variable. 1: yes, 0: no
	VAR9	Whether there is back pay behavior	virtual variable. 1: yes, 0: no
	VAR10	Whether the contract and trustworthy enterprise	virtual variable. 1: yes, 0: no
	VAR11	Whether there is owing taxes behavior	virtual variable. 1: yes, 0: no
social responsibility	VAR12	Whether to carry out energy conservation, comprehensive utilization of resources	virtual variable. 1: yes, 0: no
	VAR13	Whether the environmental behavior rating is excellent or good	virtual variable. 1: yes, 0: no

3.3. Logit regression analysis

3.3.1Model specification

Traditional methods, such as multivariate discriminant analysis method and probit method, is more stringent for data requirements, it need more assumptions. Logit regression assumptions are more relaxed, without requirements of data distribution. As the enterprise data quality is not high and the lack of data, the traditional research method is easy to cause the result error increased, so Logit model is more

suitable to study risk assessment of Dishonest enterprise enforcement.

In the Logit model, dishonest probability is regarded as a virtual variable problem virtual variable is a variable value to 0 or 1.In this model, the greater the probability calculation indicates that the greater the possibility of dishonesty. We take 0.5 as the critical value of the probability of dishonesty. If dishonesty probability is greater than 0.5, the sample will be judged to be dishonest, otherwise it will be honest.

Logit model assumes that the probability of the occurrence of the variable and the various factors that affect the following nonlinear relationship : $P = \frac{e^y}{1+e^y} = \frac{1}{1+e^{-y}}$, $y = c_0 + \sum_{i=1}^{p} c_i x_i$, x_i represents the first i explanatory variable, the c_i is the coefficient of the first explanatory variable, and P is the dependent variable, and $P \in (0,1)$.

3.3.2Preliminary modeling

1. Multicollinearity analysis

Before the Logit regression analysis, it is necessary to have correlation analysis of the 13 explanatory variables. We select the software IBM SPSS Statistic 22.Firstlythe selection of the independent variable, there was a great influence on the Logit regression when there was a severe multivariate co linearity between the variables, so we need to test the multicollinearity of 13 explanatory variables. SPSS 22 multicollinearity analysis of test results are as follows:

Table2 Multicollinearity test results

model	Non-standardized coefficients		Standardized coefficients	t	Sig	Collinearity statistics	
	B	Standard error	Beta			Tolerance	VIF
constant	.006	.000		12.974	.000		
VAR1	.003	.002	.008	1.747	.081	.980	1.021
VAR2	.002	.006	.002	.335	.738	.821	1.217
VAR3	-.009	.011	-.004	-.777	.437	.876	1.141
VAR4	-.007	.009	-.003	-.760	.447	.989	1.011
VAR5	5.442E-7	.000	.008	1.755	.079	.993	1.007
VAR6	.000	.005	.000	.050	.960	.970	1.031
VAR7	.311	.006	.247	55.107	.000	.975	1.026
VAR8	.011	.007	.008	1.628	.104	.833	1.200
VAR9	.019	.002	.039	8.591	.000	.968	1.033
VAR10	-.007	.004	-.008	-1.852	.064	.962	1.040
VAR11	.084	.003	.109	24.515	.000	.992	1.008
VAR12	.012	.009	.006	1.321	.187	.883	1.132
VAR13	-.002	.004	-.003	-.549	.583	.881	1.135

When the result of the VIF values is greater than or equal 10,that means there are serious multicollinearity between independent variables and the rest of independent variables, and this multicollinearity might undue influence least squares estimates. From table2, the selected variables corresponding to the variance inflation factor VIF is less than 10, that means there is no multivariate collinearity between variables. Therefore, we can choose these 13 explanatory variables as the indicators to establish the Logit model.

2. Variable selection

First, we set the 13 explanatory variables in the sample set to Logit regression of the dependent variable, the goodness of fit is 0.155, P (sig.) =0.000. That means the overall model is significant. All explanatory variables effectively explain the dependent variable. Then, the variables are gradually screened by the forward method significant threshold is set to 0.05, the results shows that retain the following three explanatory variables: VAR7 (whether there is a non normal credit), VAR9 (Whether there is back pay behavior)and VAR11 (Whether there is owing taxes behavior).Therefore, the final Logit regression model includes three explanatory variables as shown in Table 3:

Table3 Explanatory variables in logit regression model

measurefactor	variable code	variable name	definition method
credit information	VAR7	Whether there is a non normal credit	virtual variable. 1: yes, 0: no
credit situation	VAR9	Whether there is back pay behavior	virtual variable. 1: yes, 0: no
	VAR11	Whether there is owing taxes behavior	virtual variable. 1: yes, 0: no

3.3.3Logit regression analysis

With the use of explanatory variables VAR7, VAR9, VAR11 for Logit model, we have the following results:

Table 4 variables in equations

	B	S.E.	Wald	df	Sig.	Exp(B)
constant	-4.505	.044	10276.966	1	.000	.011

From table 4, we can see the initial assignment method of the system to the model, at the beginning, only the constant term assignment, results for B = -4.505, Standard error S.E. = 0.044, df = 1, P = Sig. = 0.000, Exp (B) = 0.011, it reached a significant level.

Table 5 comprehensive test of model coefficient

		chi-square	df	Sig.
Step1	step	833.010	3	.000
	block	833.010	3	.000
	model	833.010	3	.000

Significance level of 0.05 chi-square critical value is 7.81, in the significance test of the model, Chi square is 833.01 greater than 7.81, the corresponding Sig. value is less than 0.05, so in the case of a significant level 0.05,all are through the inspection.

Table 6 Model summary

Step	2 log likelihood	Cox & Snell R^2	NagelkerkeR^2.
Step2	4809.644	.018	.155

Statistic Cox & Snell R^2 = 0.018，Nagelkerke R^2, the square of the maximum likelihood value is used to test the integrity of the model fitting effect, the value in the theory will apply with the chi-square distribution, and the value of the table 6 is greater than the critical value given by the preceding. Thus, the maximum likelihood is passed through the numerical test, showing that the model fit well.

Table 7 variables in equations

		B	S.E.	Wald	df	Sig.	Exp(B)
Step3	VAR7	3.651	.141	672.140	1	.000	38.506
	VAR9	.983	.137	51.149	1	.000	2.673
	VAR11	2.240	.136	270.103	1	.000	9.397
	constant	-4.949	.056	7887.248	1	.000	.007

From table 7,Wald test of three explanatory variable coefficients are significant at the 0.05 level by testing, even if the significance level is 0.01, it also can pass the test. It shows that the three explanatory variables have good explanatory power to the model. The above test results show that the Logit model constructed with these three explanatory variables is significant. The above results can help to predict the risk of enterprise dishonest enforcement.

The resulting Logit model:

$$\ln\frac{P}{1-P} = 3.651VAR7 + 0.983VAR9 + 2.24VAR11 - 4.949$$

Or:

$$P = \frac{\exp(3.651VAR7 + 0.983VAR9 + 2.24VAR11 - 4.949)}{1 + \exp(3.651VAR7 + 0.983VAR9 + 2.24VAR11 - 4.949)}$$

3.3.4 Model test

Table 8effectiveness of the Logit model

	observed	predicted		
		virtual variables of dishonest to enforcement		percentage Correction
		0	1	
Step4	no dishonest to enforcement	0	46296	19
	dishonest to enforcement	1	482	30
	total percentage			98.9

From table 8:Logit regression model is almost 100%accuracy (99.959%) to predict the no behavior of dishonest enterprise to enforcement, the correct rate to exist the behavior of dishonesty is 5.9%, the overall prediction accuracy is 98.9%.

Although the overall prediction accuracy of the model is very high, but the recognition accuracy rate is not high for the dishonest behavior enterprise executed, only 5.9%. However, we notice the proportion of dishonest enterprise sample accounts for the proportion of the total sample is only 1.09%, the recognition accuracy, compared to random probability, have improved a lot. For those honesty and executed acts of enterprises, its recognition accuracy rate is almost 100 percent. So it has application value.

In addition, after screening the rest of three explanatory variables are highly significant. It shows that these three indicators are very effective for explaining the risk of losing the credit of the enterprise. Only the factors of the data set are not enough, causes the explanatory power of model is not enough for the risk of dishonest enterprise to enforcement.$R^2$is only 15.5%. Finally, the screening of the remaining three variable VAR7 (Whether there is a non normal credit), VAR9 (Whether there is back pay behavior) and VAR11 (Whether there is owing taxes behavior) is about the enterprise credit behavior. That means bad credit behavior indicates that the future of dishonesty behavior,

4. Model application

4.1Background

Based on Logit model, the risk assessment of dishonest cooperate executed, has been a preliminary experimental application in C city, J province in China. As part of the credit risk assessment, the model features embedded in the city's credit system.

The city has a total population of 2 million including residential and transient, and a gross land area of 1,500 square kilometers.

The credit system, constructed by city XX, has a portal site window 'integrity XX', four central credit information databases, a supporting public credit information system and key development products of credit information service. Throughout a continuously two-year constructing and applying, the city-wide society credit framework has been constructed.

Enterprise credit information database, Person credit information database, Financial credit information database, Government credit information database. The unified and standardized credit information databases have been already built. It includes enterprise credit information database, person credit information database, financial credit information database and government credit information database. The cumulative storage has covered the public credit information of 49 departments. Specifically, it contains 221 types of information, 2491 items of information and a total amount of over 12,500,000 data (including 6,790,000 business data and 5,710,000 basic data).

For C city's credit system development environment server, Windows Server 2003 is adopted for development environment on the server end, IIS WEB internet information server for operating system, Oracle 11g is as the database and JAVA as the development tool.

4.2Effects

Before experimental model application, Ccity sums up Corporate Credit Risk to achieve Credit risk assessment, by various government departments. The

departments build different business database models. The assessment models are also different, including assessment caliber and the accuracy rate.

Based on Logit model, the risk assessment of dishonest cooperate executed, has been a preliminary experimental application in C city, J province in China. It shows some perspectives in the following. First, modeling data is from various commissioned Bureau of comprehensive data, rather than a single commissioned Bureau; Second, The model removes unrelated variables to dishonest executed risk; Third, The assessment results are basically the same as the original methods; Finally, The assessment results are corresponding to each enterprise.

And follow-up, we will increase iterations to the observed effects as data increasing.

5. Conclusions and future work

The Smart City's construction emphasizes building an effective whole social credit system, promoting the construction of government integrity, business integrity, social integrity and public confidence in the judiciary. Wherein, the assessment of corporate credit risk is very significant for urban development, market management, social management, and residents' life, etc. As the construction of credit system, one of the three major data administration credit system need to involve all levels of government credit operations break down departmental boundaries, to achieve interoperability of data. The establishment of administration credit, based on e-government, is also the one of the main content of Chinese Thirteen Five construction of smart city.

In this paper, we construct the modeling of Corporate Credit risk executed, on the basis of governmental big data mining. The use of enterprise data set of Chinese C city of J Province, choose the appropriate explanatory variables from five aspects, administrative punishment, innovation, credit information, credit situation, and social responsibility, to screen corporate dishonesty behavior Logit regression, to find out significant predicted variables for corporate Credit risk executed, and finally construct a Logit regression model from these variables filtered out. In the model test, we find that, although the overall prediction accuracy of the model is high, but the recognition accuracy, for the existent dishonest behavior for executed corporate, is not high.

However, it is noted that the sample proportion of dishonesty and the overall is very small, the recognition accuracy still improve a lot, more than random probability. And the recognition accuracy rate is almost 100%, for the honesty corporate with executed behavior, its. So it's valuable applying. The successful application of Logit model for the risk assessment of dishonest corporate executed, makes the foundation for further integrated assessment of the credit risk business. In addition, the significant variable in the model is on business credit behavior, indicating that bad credit behavior is closely linked, between each other. It also reveals the necessity to build urban enterprises comprehensive credit data set, in the context of urban construction of Smart City.

Three working aspects in the future. First, we will increas modeling analysis on unstructured information, on the basis of current data; Second, we will modeling analysis on Internet information, on the basis of current data; Thirdly, we will deepen the application in C city, J province, China, following and observing the internal regulations of the dishonest corporate executed risk behavior. Finally, we will expand more risk assessment study for more Chinese cities and more fields.

Acknowledgments

The authors would like to thank Qili Wang at School of Information, Renmin University of China to help process enterprise credit information, and also thank the anonymous reviewers for their valuable comments and suggestions to this paper.

References

[1] M. Corazza, S. Funari, R. Gusso, "An evolutionary approach to preference disaggregation in a MURAME-based creditworthiness problem", Applied Soft Computing, 2015, vol. 29, pp. 110–121.

[2] R, Tsaiha, Y.J. Liu, W. Liu, Y.L. Lien, "Credit scoring system for small business loans", Decision Support Systems, 2004, vol. 38, pp. 91-99.

[3] G. Andreeva, R. Calabrese, S. Osmetti, "A comparative analysis of the UK and Italian small businesses using generalized extreme value models", European Journal of Operational Research, 2016, vol. 249, pp. 506-516.

[4] C. Serrano-Cinca, B. Gutierrez-Nieto, N. M. Reyes, "A social and environmental approach to microfinance credit scoring", Journal of Cleaner Production, 2015, vol. 112, pp. 3504-3513.

[5] G.B. Fernandes, R. Artes, "Spatial dependence in credit risk and its improvement in credit scoring", European Journal of Operational Research, 2016, vol. 249, pp. 517-524.

[6] Y. Ju, S.Y. Jeon, S.Y. Sohn, "Behavioral technology credit scoring model with time-dependent covariates for stress test", European Journal of Operational Research, 2015, vol. 242, pp. 910-919.

[7] D. Zhang, W. Xu, Y. Zhu, X. Zhang, "Can sentiment analysis help mimic decision-making process of loan granting? A novel credit risk evaluation approach using

GMKL model", The 48th Hawaii International Conference on System Science (HICSS), 2015, pp. 949-958.

[8] P. Kolesar, J.L. Showers, "A robust credit screening model using categorical data", Management Science, 1985, vol. 31, no. 2, pp. 124-133.

[9] C.L. Huang, M.C. Chen, and C.J. Wang, "Credit scoring with a data mining approach based on support vector machine", Expert Systems with Applications, 2007, vol. 33, no. 4, pp. 847-856.

[10] H.L Jensen, "Using neural networks for credit scoring", Managerial Finance, 1992, vol. 18, no. 6, pp. 15-26.

[11] V.S. Desai, D.G. Conway, J.N. Crook, and G.A. Overstreet, "Credit-scoring models in the credit-union environment using neural networks and genetic algorithms", IMA Journal of Management Mathematics, 1997, vol. 8, no. 4, pp. 323-346.

[12] K.B. Schebesch, R. Stecking, "Support vector machines for classifying and describing credit applicants: detecting typical and critical regions", Journal of the Operational Research Society, 2005, vol. 56, no. 9, pp. 1082-1088.

[13] A. Khashman, "A neural network model for credit risk evaluation", International Journal of Neural Systems, 2009, vol. 19, no. 04, pp. 285-294.

[14] D. West, "Neural network credit scoring models", Computers & Operations Research, 2000, vol. 27, no. 11, pp. 1131-1152.